Environmental and Ecological Statistics **8,** 171–180, 2001

Contributions to composite sampling

GERALD VAN BELLE, WILLIAM C. GRIFFITH and
STEVEN D. EDLAND

Department of Environmental Health, School of Public Health and Community Medicine, University of Washington, Seattle WA 98195

Received September 1997; Revised November 2000

The initial use of composite sampling involved the analysis of many negative samples with relatively high laboratory cost (Dorfman sampling). We propose a method of double compositing and compare its efficiency with Dorfman sampling. The variability of composite measurement samples has environmental interest (hot spots). The precision of these estimates depends on the kurtosis of the distribution; leptokurtic distributions $(\gamma_2 > 0)$ have increased precision as the number of field samples is increased. The opposite effect is obtained for platykurtic distributions. In the lognormal case, coverage probabilities are reasonable for $\sigma < 0.5$. The Poisson distribution can be associated with temporal compositing, of particular interest where radioactive measurements are taken. Sample size considerations indicate that the total sampling effort is directly proportional to the length of time sampled. If there is background radiation then increasing levels of this radiation require larger sample sizes to detect the same difference in radiation.

Keywords: Dorfman sampling, double compositing, Poisson and lognormal compositing, leptokurtic and platykurtic distributions.

1. Introduction

Composite sampling goes back to the Second World War when blood of army recruits was pooled in the process of determining the presence of the syphilis bacterium (Dorfman, 1943). The rarer the occurrence of syphilis the more efficient the composite sampling strategy. Since the introduction of this strategy a host of applications have been made, including scenarios where the amount of a substance is to be measured—not just the presence or absence. A comprehensive, annotated bibliography summarizing papers on composite sampling from 1936 to 1992 can be found in Boswell *et al.* (1996). This paper makes a useful distinction between sampling for presence or absence of a characteristic and estimation. In the latter case there is additional emphasis on variance of the estimate. Closely related to composite sampling are double sampling and ranked set sampling. An annotated bibliography on ranked set sampling can be found in Kaur *et al.* (1995). An informative discussion of, and application to, bioremediation can be found in O'Brien and Gilbert (1997). Other examples of composite sampling have involved determining leaks in sealed radioactive sources (Thomas *et al.*, 1973), searching for (rare) high grade ore (Garrett and Sinding-Larson, 1984), ground water screening (Rajagopol and Williams, 1989), and assessing the effectiveness of clean-up of PCB's (listed in Boswell *et al.*, 1996).

While composite sampling is not a panacea for environmental estimation it does have potential for substantial saving in sampling costs. One barrier to greater use of composite sampling has been regulatory requirements that are usually couched in terms of single field samples and do not allow pooling of samples. It will take specific demonstrations by statisticians to show that in certain situations regulatory requirements can be met with composite sampling.

A second reason why composite sampling is less popular than it could, or should be, is that chemists have a bias against false positives so that detection levels are frequently made much higher than necessary. It is better to say, ''below the detection level'' than to claim that the substance is not present.

2. Dorfman sampling

Following Dorfman (1943), let π be the prevalence rate of positive field samples, n the number of field samples in a composite sample, π^* the probability that the composite sample is positive, N the number of field samples to be analyzed (so that there are N/n composite samples). Let T be the total number of composite samples that have to be analyzed in detail. Then the expected value, $E(T)$, is,

$$E(T) = \frac{N}{n} + n\left(\frac{N}{n}\right)\pi^*. \tag{1}$$

The ratio of the expected number of composite samples to the total number of field samples is a measure of the relative cost of composite sampling to simple sampling,

$$\text{Relative Cost} = \frac{E(T)}{N} = \frac{1}{n} + \pi^* = \frac{n+1}{n} - (1-\pi)^n. \tag{2}$$

This function can be minimized for n relative to the prevalence π or *vice versa*.

The rarer this characteristic the lower the relative cost. The Relative Cost curve relative to composite pool size, n, is rather flat for low prevalences. For example, for $\pi = 0.001$ the values the Relative Cost for $n = 20, 30, 50$ are 0.070, 0.063 and 0.069, respectively. Fig. 1 plots the relative cost as a function of the prevalence, π. If there is uncertainty about the value of the prevalence this figure indicates that there is little advantage in taking very large composite samples for the purpose of detecting the presence or absence of a characteristic. The figure suggests that as a rule of thumb a composite pool size of about five yields substantial savings over a large range of prevalences.

3. Double compositing

We now describe a technique which enhances the compositing approach. Where compositing of samples is done rather easily, such as collecting water at a tap for lead analysis, this approach has specific advantages as will be illustrated. The proposed scheme is called double compositing and involves collecting two (or more) composite samples from a source. The basic idea is to take double samples and group them in a grid pattern; this could be done by physically collecting two field samples at one location or taking field

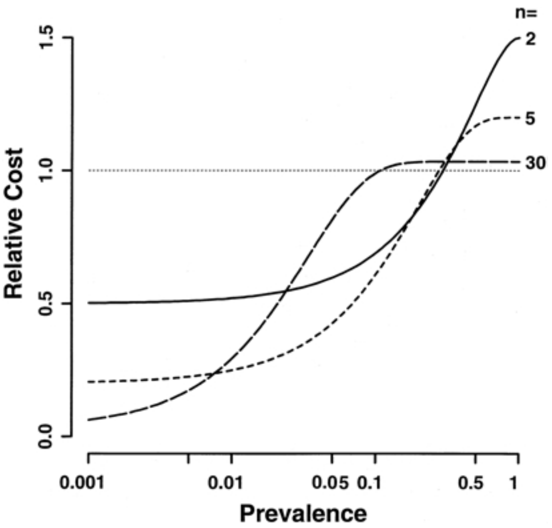

Figure 1. Cost effectiveness of composite sampling as function of prevalence for three composite sample sizes.

samples and splitting them in the laboratory. One obvious advantage is that if the characteristic looked for is rare, then by this technique the sample is immediately identified if only one of the composited rows and columns is positive. Also if more than one row and column is positive then only the samples at the intersection of the positive rows and columns have to be retested. This approach may be particularly fruitful in spatially correlated observations such as hot spots.

Let K and L be the number of rows and columns for the double compositing scheme, then the rows can be thought of as K single composite samples of length L and the columns as L single composite samples of length K. From the results for single composite samples the probability of at least one positive sample in a row is $\pi_L^* = 1 - (1 - \pi)^L$ and we expect $\pi_L^* K$ of the row composite samples to be positive. Similarly we expect $\pi_K^* L$ of the column composite samples to be positive. To compute $E(T)$, the expected number of samples analyzed, a recursive formula has been developed for the conditional probability of k positive rows and l positive columns given that m is the number of true positive field samples. The recursive formula is:

$$\begin{aligned}[KL - (m-1)]P[k,l|m] &= P[k,l|m-1][kl-(m-1)]\\ &\quad + P[k-1,l|m-1]l[K-(k-1)]\\ &\quad + P[k,l-1|m-1]k[L-(l-1)]\\ &\quad + P[k-1,l-1|m-1][K-(k-1)][L-(l-1)].\end{aligned}$$

The probabilities are zero if $k > m$, $l > m$, $k = 0$, $l = 0$, $m = 0$, $kl < m$, $k > K$, $l > L$, and the initial condition is, $P[1,1|1] = 1$.

The relative cost of this scheme is the expected number of tests, $E(T)$, as before divided by the number of tests, KL, without compositing as illustrated in Fig. 2 for the case where $K = L$. The relative costs for double composite samples are lower than single composite

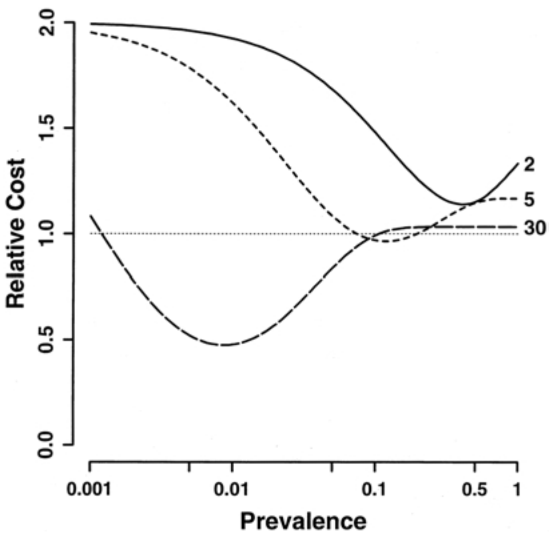

Figure 2. Cost effectiveness of double compositing compared with Dorfman compositing.

samples for a range of low prevalences and large pool size. Single composite samples have a lower cost for both high prevalences and very low prevalences for a given pool size. The lower cost at very low prevalences for single composite samples occurs approximately at prevalences where the expected number of positive samples in the *KL* samples is less than 1. Double composite samples can significantly reduce costs compared to single composite samples over a range of low prevalences provided a sufficiently large pool sizes can be used. Also as the pool size increases the costs for double composite samples decrease over a wider range of prevalences. Fig. 2 also indicates that as the prevalence becomes small the pool size becomes large, but the curve is very flat indicating the insensitivity of relative cost to pool size.

A minor improvement can be made if an overall composited sample can be analyzed first. This would be equivalent to Dorfman sampling. If the overall sample were positive the row and column scheme proposed here would be applied. This would require three separate instances of assays and may not be practicable.

4. Estimating variability

In composite measurement sampling field samples are composited with a view to reducing laboratory analysis while maintaining or increasing the precision of the estimates of concentration. We will follow the notation of Edland and van Belle (1994). The samples to be pooled are called the field samples and values of the endpoint of interest will be denoted by the random variable X. The samples to be composited and analyzed are called the laboratory samples and endpoints denoted by Y. Unless indicated otherwise we will assume the n field samples are composited into one laboratory sample and I laboratory samples are analyzed. Specifically, the i-th laboratory sample, Y_i is defined by $Y_i = \frac{1}{n}\Sigma X_{ij}$.

Let the mean and variance of the field sample, X, be μ_x and σ_x^2, respectively. Suppose that $X_{ij}, i = 1, 2, \ldots, I, j = 1, 2, \ldots, n$ are independent, identically distributed samples. Let the mean and variance of the composited samples be μ_y and σ_y^2. Then $\sigma_x^2 = n\sigma_y^2$. Basic results associated with this formulation can be found in Lovison (1993) and Edland and van Belle (1994).

In environmental sampling a key parameter of interest is the variability, rather than the mean. For example, the search for ''hot spots'' deals with the variability of the site, not the mean level of contamination. Hence, we are particularly interested in the precision of $\hat{\sigma}_x^2$. The precision of this estimate of the variance depends on the fourth order moments of the distribution being sampled. The kurtosis, γ_2, for a random variable is

$$\gamma_2 = \frac{\mu_4}{(\sigma^2)^2} - 3,$$

where μ_4 is the fourth central moment and σ^2 is the variance. The variance of the estimate of the variance of the composited sample is

$$V_{\gamma_2}(\hat{\sigma}_x^2) = n^2 \frac{2(\sigma_y^2)^2}{I-1}\left[1 + \frac{(I-1)}{nI}\frac{\gamma_2(Y)}{2}\right], \tag{3}$$

which, in terms of the distribution of the field samples becomes

$$= \frac{2}{I-1}(\sigma_x^2)^2 + \frac{\gamma_2(X)}{I\ n}(\sigma_x^2)^2. \tag{4}$$

Table 1 lists the values of the kurtosis statistic for a variety of distributions. For the normal case $\gamma_2 = 0$ and the right hand term drops out. In this case, the precision of the estimate of the variability of X is only a function of the number of composite samples, I, that are taken. Distributions with $\gamma_2 > 0$ (leptokurtic) are characterized by improved precision of the estimate of the variability as the number of field samples in the composite sample increase. For $\gamma_2 < 0$ (platykurtic) this precision decreases since the term on the right hand side becomes smaller as the number of field samples (or composite samples)

Table 1. Kurtosis for a variety of distributions.

Distribution of X	$\gamma_2(X)$	*Comments*
Normal	0	
Poisson	$\frac{1}{\lambda}$	Approaches 0 as $\lambda \to \infty$
Uniform (continuous)	-1.2	Independent of range
Uniform (two-point)	-2.0	''Worst case''
Unimodal symmetric	> -1.2	
Double exponential	3	
Exponential	6	
t-distribution	$\frac{6}{(\nu-4)}$	$\nu > 4$
Lognormal	$(\omega - 1)(\omega^3 + 3\omega^2 + 6\omega + 6)$	where $\omega = e^{\sigma^2}$
$\sigma^2 = 1$	112.29	
$\sigma^2 = 0.1$	1.856	
Binomial	$\frac{(1-6pq)}{npq}$	
$0.2113 < p < 0.7887$	< 0	
$p < 0.2113, p > 0.7887$	> 0	

increase. The ''worst case'' scenario (where compositing decreases the precision of the estimate of variability) is the two point uniform distribution. In this case $\gamma_2 = -2.0$. The uniform distribution has $\gamma_2 = -1.2$ and this is the smallest value that can be taken on by any symmetric distribution (Kendall, Stuart and Ord, 1987, page 107).

Table 1 also indicates that in the Poisson case (and the binomial) the value of γ_2 approaches zero as the mean increases. This, of course, is consistent with the fact that the distributions approach normality as the mean increases. Thus for large values of the mean of these distributions we can assume the normal model.

The binomial distribution is platykurtic for values of $0.2113 < p < 0.7887$ and leptokurtic for values outside this range. Hence, when samples are composited with conditions that occur rarely (or most of the time) the precision of the estimate of variability is increased with compositing.

5. Large sample lognormal case

The lognormal distribution is of considerable environmental interest as the most common distribution after the normal. Little work has been done on implementing composite sampling for the lognormal distribution. From the previous section some conclusions can be drawn about the lognormal. Table 1 indicates that the second term in Equation (4) can be very important when there is considerable variability. For a variance of 1, $\gamma_2 = 112.29$, and the second term dominates the first term unless n is made very large. For a variance of 0.1 the second term is much less dominant. The conclusion from this analysis is that in the lognormal case with substantial field variability there is considerable benefit in compositing in order to increase the precision of the estimate of variability.

Since an increase in n makes the central limit theorem applicable we investigated coverage probabilities of lognormally composited samples. A first approach uses the central limit theorem. Since the total sample size to be considered is nI it is reasonable to investigate the accuracy of the lognormal distribution. Table 2 lists the coverage of confidence intervals for samples from a lognormal distributions with standard deviations ranging from 0.1 to 2.0. For standard deviations of 0.5 or less the coverage probabilities are reasonably good. An analytic solution is required for samples from more extreme distributions. This is consistent with the findings of Barakat (1976) who discusses the ''permanence of the lognormal probability density function'' for sums of lognormal random variables.

Table 2. True coverage probabilities for nomimal 95% confidence intervals when sampling from lognormally distributed field samples. Monte Carlo sampling of 50,000 samples.

	σ			
Sample Size	0.1	0.5	1.0	2.0
$nI = 5$	0.95	0.93	0.86	0.61
$nI = 20$	0.95	0.93	0.89	0.67
$nI = 50$	0.95	0.94	0.92	0.74
$nI = 100$	0.95	0.95	0.93	0.78
$nI = 200$	0.95	0.95	0.94	0.82

6. Poisson case for estimating radiation counts above background

The Poisson distribution is important for modeling radioactive counts and composite sampling has potential application. This application is an example where quantitative estimation is of interest, not just presence or absence of a characteristic. A feature that is particularly attractive is that the radioactivity of composited field samples is the sum of the radioactivity of the individual field samples.

We will work with the total of composited field samples rather than the mean. Assume that X_{ij} are i.i.d. Poisson(θ). Let $V_i = \Sigma X_{ij}$ be the total for the n field samples making up the composite. Then $V_i \sim$ Poisson$(n\theta)$. Let $W = \Sigma V_i$ be the sum of I composite samples. Then W is Poisson$(nI\theta)$. Using the square root transformation as a variance stabilizing and normalizing transformation we get, in units of a field sample that, asymptotically:

$$Z = \sqrt{\frac{W}{nI}} \sim N\left(\sqrt{\theta}, \frac{0.25}{nI}\right).$$

This result is particularly simple and allows the immediate application of all the results for the normal case found in Edland and van Belle (1994). We illustrate in the case of comparison of two independent samples. The usual sample size calculation formula for this situation (Fisher and van Belle, 1993) is

$$nI = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2 \sigma^2}{(\mu_1 - \mu_2)^2}. \tag{5}$$

In many sample size calculations the Type I error is fixed at 0.05, and the power is required to be 0.80. For $\alpha = 0.05$, $\beta = 0.20$ the values of $z_{1-\alpha/2}$ and $z_{1-\beta}$ are 1.96 and 0.84 respectively and $2(z_{1-\alpha/2} + z_{1-\beta})^2 = 15.68$ which can be rounded up to 16. So a quick rule of thumb for sample size calculations is

$$nI = \frac{16\sigma^2}{(\mu_1 - \mu_2)^2}.$$

In the Poisson case using the square root transformation and n field samples per composite sample, and I composite samples, the formula becomes

$$nI = \frac{4}{(\sqrt{\theta_1} - \sqrt{\theta_2})^2}.$$

For example, if the mean radioactivity is assumed to be 0.5 counts per minute under one condition and 0.25 counts per minute under another condition then the sample size required per group is 23. This total of 23 observations per group can be obtained by either using twenty three field samples or, for example, compositing field samples in units of five and analyzing five composite samples.

In the Poisson case the means are frequently a function of the length of time observed such as radioactive counts per unit time. This is an example of composite sampling in time rather than space. The composite sampling strategy can then be formulated explicitly in terms of the length of time that the field samples are counted. Let the means θ_1 and θ_2 be the means per unit time and the observation period per field sample of duration, T. Then V_i are Poisson with mean $T\theta_i$. Hence the sample size required can be shown to be

$$nI = \frac{4}{T(\sqrt{\theta_1} - \sqrt{\theta_2})^2}.$$

By increasing the observation period T we reduce the sampling effort proportionately to T, not as the square root of T. For example, doubling the observation period of the field sample reduces the required number of laboratory samples by a factor of two. Similarly, doubling the laboratory observation time reduces the number of field samples to be taken by the same factor. Thus there is an interplay between field samples, laboratory samples and observation time that can be explored for specific situations.

A test for Poissonness can be carried out on the composite sample totals. Let $\overline{V}$ and s_v^2 be the sample mean and variance of the composite samples. Then under the assumption that the composite sample means come from a Poisson distribution the quantity,

$$\frac{(I-1)s_v^2}{\overline{V}}$$

is approximately chi-square with $I - 1$ degrees of freedom. This is the usual Poisson homogeneity test. Large values of the statistic indicate over-dispersion and small values under-dispersion.

We now discuss the case of Poisson data with background radiation. Suppose that the background level of radiation is θ^* and let θ_1 and θ_2 be the additional radiation over background. Then, X_i is Poisson $(\theta^* + \theta_i)$. We can then apply the sample size formula in terms of square roots. However, this can be more transparently approximated by the usual sample size formula in the original scale and assuming the normal approximation to the Poisson distribution. The rule-of-thumb sample size formula is

$$nI = \frac{16[\theta^* + (\theta_1 + \theta_2)/2]}{(\theta_1 - \theta_2)^2}.$$

where the variance of the response in the two populations is estimated by $\theta^* + (\theta_1 + \theta_2)/2$. The denominator does not include the background radiation but the numerator does. Since the sample size is proportional to the numerator, increasing levels of background radiation require larger sample sizes to detect the same difference in radiation levels.

Table 3 considers sample sizes for a variety of background levels. The first row of the table considers the two sample situation when there is no background radiation and the means are 1 and 2. The total sampling effort is $nI = 24$, which can be split, again, between field and laboratory efforts. As the background increases the sample sizes increase. For example, when the background is 1.5 (and thus constitutes sixty percent of the first sample mean) the required sample size is 48, double the sampling effort at no background radiation.

Table 3. Sample size and background radiation. Two sample case: n is the number of field samples per I composite samples. Numbers in parentheses in last column are sampling effort as percent of effort when there is no background radiation ($\theta^* = 0$).

Background θ^*	*Mean Sample 1* $\theta^* + \theta_1$	*Background %*	*Mean Sample 2* $\theta^* + \theta_2$	nI
0	1	0	2	24.0
0.1	1.1	9	2.1	24.8 (1.03)
0.5	1.5	33	2.5	32.0 (1.33)
1.0	2.0	50	3.0	40.0 (1.67)
1.5	2.5	60	3.5	48.0 (2.00)
5.0	6.0	83	7.0	104.0 (4.33)

7. Discussion

Composite sampling has potential for both application and development of new methodological approaches. Researchers should become aware of the potential of this, and other, statistical techniques that may save substantial sums of money.

The Poisson model with background noise is a useful model for epidemiological investigations of effects of environmental pollution, for example, investigating whether disease occurrence is above background level.

Creative adaptations of compositing can be developed for specific situations. For example, if a cheap measurement can be made of a variable of interest and this measurement is highly correlated with a more expensive measurement it may be cost effective to composite and measure the cheap variable first. If it has low value then no further testing would be done.

An area that requires further research is power calculations for lognormal field sample distributions

Acknowledgments

This research was supported by the Consortium for Risk Evaluation with Stakeholder Participation (CRESP) by Department of Energy Cooperative Agreement #DOE-FCO1-95EW55084 and by an EPA Grant. Although the research described in this article has been funded in part by the United States Environmental Protection Agency through agreement CR825173-01-0 to the University of Washington, it has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred.

References

Barakat, R. (1976) Sums of independent lognormally distributed random variables. *Journal of the Optical Society of America*, **66**, 211–16.

Boswell, M.T., Gore, S.D., Lovison, G., and Patil, G.P. (1996) Annotated bibliography of composite sampling Part A: 1936-92. *Environmental and Ecological Statistics*, **3**, 1–50.

Dorfman, R. (1943) The detection of defective members of a large population. *Annals of Mathematical Statistics*, **14**(4), 436–40.

Edland, S.D. and van Belle, G. (1994) Decreased sampling costs and improved accuracy with composite sampling. In: C.R. Cothern and N. P. Ross (eds) *Environmental Statistics, Assessment, and Forecasting*, Lewis Publishers, pp. 29–55.

Fisher, L.D. and van Belle, G. (1993) *Biostatistics: A Methodology for the Health Sciences*. Wiley and Sons, New York.

Garrett, R.G. and Sinding-Larsen, R. (1984) Optimal composite sample size selection, applications in geochemistry, and remote sensing. *Journal of Geochemical Exploration*, **21**, 421–35.

Kaur, A., Patil G.P., Sinha, A.D., and Taillie, C. (1995). Ranked set sampling: an annotated bibliography. *Environmental and Ecological Statistics*, **2**, 25–54.

Kendall, M.G., Stuart, A., and Ord, J.K. (1987) *Kendall's Advanced Theory of Statistics. Volume 1: Distribution Theory*. Fifth edition. Oxford University Press, New York, p. 107.

Lovison, G. (1993) A unified linear model estimation with composite sample data. In G.P. Patil and C.R. Rao, (eds) *Multivariate Environmental Statistics*, North Holland Series in Statistics and Probability, Volume 6, 255–87.

O'Brien, R.F. and Gilbert, R.O. (1997). Sample designs using in situ measurements. In *In Situ and On-Site Bioremediation. Volume 5*. B.C. Alleman and A. Leeson (Symposium Chairs), Battelle Press, pp. 357–62.

Rajagopal, R. and Williams, L.R. (1989) Economics of sample compositing as a screening tool in ground water quality monitoring. *Ground Water Monitoring Review*, **9**, 186–92.

Thomas, J., Pasternack, B.S., Vacirca, S.J., and Thompson, D.L. (1973) Application of group testing procedures in radiological health. *Health Physics*, **25**, 259–66.

Biographical sketches

Gerald van Belle is professor in the Departments of Biostatistics and Environmental Health at the University of Washington.

William C. Griffith is a staff member in the Department of Environmental Health at the University of Washington.

Steven D. Edland is a staff member at the Mayo Clinic.